UNIVERSIDADE FEDERAL DE SANTA CATARINA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Rafael Dobrachinsky Plentz

REDES BAYESIANAS PARA ANÁLISE DE COMPORTAMENTO APLICADAS A TELEFONIA CELULAR

Dissertação submetida à Universidade Federal de Santa Catarina como parte dos requisitos para a obtenção do grau de Mestre em Ciência da Computação

Dr. Carlos Becker Westphall

Florianópolis, Março de 2003

REDES BAYESIANAS PARA ANÁLISE DE COMPORTAMENTO APLICADAS A TELEFONIA CELULAR

Rafael Dobrachinsky Plentz

Esta Dissertação (Tese) foi julgada adequada para a obtenção do título de Mestre em Ciência da Computação Área de Concentração Sistemas de Computação e aprovada em sua forma final pelo Programa de Pós-Graduação em Ciência da Computação.

_____

Fernando Álvaro Ostuni Gauthier, Dr.

Coordenador do Curso

Banca Examinadora:

_____

Prof. Carlos Becker Westphall , Dr.

Orientador

_____

Prof. Marcelo Menezes Reis, Dr.

_____

Prof.ª Silvia Modesto Nassar, Drª.

Agradecimentos

Aos meus pais que me apoiaram no início desta minha caminhada e incentivo em todas as oportunidades de crescimento pessoal.

Aos meus amigos da INF, aos quais estiveram presentes no dia-a-dia de minhas pesquisas e que colaboraram nas discussões sobre os mais variados assuntos.

Ao LRG, que ofereceu condições de infra-estrutura para a realização das atividades.

Ao meu orientador Prof. Carlos Becker Westphall , Dr. que me acolheu na primeira conversa sobre possíveis pesquisas em sua área. Também ao Prof. Fernando Cruz, Dr. que esteve presente em minhas pesquisas bayesianas e disponibilizou seu tempo para discutirmos sobre vários assuntos.

Ao Edson Mello que contribuiu na liberação de alguns serviços de rede para testes e simulações em ambiente seguro.

Às funcionárias da PGCC , em especial a Vera, que sempre foram prestativas no atendimento e solução de dúvidas.

À UFSC que proporcionou as condições necessárias para a realização deste trabalho.

Resumo

O aprendizado automático em redes Bayesianas faz uso do Teorema de Bayes que é de grande importância para o cálculo de probabilidades. A teoria de probabilidade envolve métodos de propagação de crença e métodos para aprendizado destas redes, focando principalmente, o aprendizado através da inferência lógica, visto que o mesmo pode ser entendido como a base para analisar um conjunto de informações disponíveis e chegar a uma conclusão objetiva, expressa numericamente.

A busca soluções para problemas identificados por uma mineração de dados, faz-se necessário a construção de um Sistema para Gerência em Redes Bayesianas para alterações de comportamento, quantificando as probabilidades associadas a aquisição do conhecimento pela exploração das informações em bancos de casos.

Abstract

The automatic learning in Bayesianas nets makes use of the theorem of Bayes that is of great importance for the calculation of probabilities. The probability theory involves Methods of propagation of belief and methods for learning of these nets, focando mainly, the learning through the inference logical, since the same it can be understood as the base to analyze a set of available information and to arrive at an objective conclusion, expressed numerically.

The search solutions for problems identified for a mining of data, becomes necessary the construction of a model for management in Bayesianas Nets for behavior alterations, quantifying the probabilities associates the acquisition of the knowledge for the exploration of the information in cases of banks.

Sumário

CAPÍTULO I..... 11

1. INTRODUÇÃO 11

1.1. OBJETIVOS..... 13

1.2. OBJETIVOS ESPECÍFICOS..... 14

1.3. ESTRUTURA DO TRABALHO 14

CAPÍTULO II..... 15

2. DATA MINING (DM)..... 15

2.1. PROCESSO DE EXTRAÇÃO DE CONHECIMENTO DE BD 17

2.2. MÉTODOS DE MINERAÇÃO DE DADOS 19

2.3. ABORDAGEM PARA MINERAÇÃO DE DADOS..... 21

2.4. A DESCOBERTA DE CONHECIMENTO EM APLICAÇÕES DE NEGÓCIOS..... 22

2.4.1. ANÁLISE DE DADOS 23

2.4.2. DESENVOLVIMENTO DE MODELOS 24

CAPÍTULO III 26

3. TEOREMA DE BAYES..... 26

3.1. INFERÊNCIA BAYESIANA 28

3.2. REDES BAYESIANAS 29

3.3. DEPENDÊNCIA E INDEPENDÊNCIA DE EVENTOS 32

3.4. PROBABILIDADES DE EVENTOS DEPEDENTES 33

3.5. INDEPENDÊNCIA CONDICIONAL 34

3.6. TRABALHOS CORRELATOS..... 35

CAPÍTULO IV **36**

4. DOMÍNIO DA APLICAÇÃO **37**

4.1. PROTÓTIPO 37

4.2. BANCO DE DADOS 37

4.3. APLICAÇÃO 39

4.4. FERRAMENTAS 42

4.4.1. NÉTICA 42

4.4.2. BAYESIAN KNOWLEDGE DISCOVERER (BKD) 44

CAPÍTULO V **46**

5. CONCLUSÕES **46**

6. TRABALHOS FUTUROS **47**

6.1. APLICAÇÃO DE TÉCNICAS DE AMOSTRAGEM DE DADOS 47

CAPÍTULO VI **48**

7. REFERÊNCIAS **48**

Lista de Abreviaturas

BBN - *bayesian belief network*

BD - Banco de Dados

BKD - Bayesyan Knowledge Discoverer

DM - *Data Mining*

DW - *Data Warehouse*

HP - *Hewlett Packard*

IA - Inteligência Artificial

KDD - *Knowledge Discovery Data*

LANs - Local Area Networks

OLAP - *Online Analytical Processing*

RAD - Rapid Application Development

RB - Redes Bayesianas

SGBD - Sistema Gerenciador de Banco de Dados

SQL - *Structure Query Language*

WANs - Wide Area Networks

Lista de Figuras

Figura 1: Análise de Probabilidades 28

Figura 2: Rede Bayesiana 30

Figura 3: Espaço Amostral 33

Figura 4: Modelagem 38

Figura 5 : Protótipo Bayesiano 40

Figura 6 :Simulação de Ligações 40

Figura 7: Nética 43

Figura 8– tela de abertura do software BKD 45

Lista de Tabelas

Tabela 1: Métricas 39

Capítulo I

1. Introdução

O gerenciamento é usado para controlar as atividades e monitorar os recursos da rede. Simplificando, o trabalho básico da gerência de rede é obter informação, extraída de dados, para um possível diagnóstico e execução de ações para resolver os problemas. Para alcançar estes objetivos, as funções de gerenciamento devem estar contidas em diversos componentes da rede permitindo o diagnóstico, a prevenção e a reação para problemas [WES96].

Esta detecção e reação aos problemas podemos chamar de gerência pró-ativa. Os modelos computacionais com comportamento inteligente são relevantes a essas reações os quais são princípios básico da Inteligência Artificial (IA). Com este pensamento, a teoria de probabilidade proporciona um relato coerente de como a crença em um dado conhecimento deve se alterar na presença de informações parciais ou incertas [PEARL88].

A necessidade de uma grande quantidade de dados para cálculos probabilísticos enumerando assim todas as relações entre as variáveis do problema, fez por muito tempo esta teoria probabilística, computacionalmente inviável e deixada de lado pelos pesquisadores.

As ferramentas de mineração são programas inteligentes ou agentes, incorporando formas de inteligência artificial em banco de dados relacional. Existem agentes determinando padrões e alterando usuários de variações. Alguns tipos de ferramentas são usados em mineração de dados, como redes neurais, árvores de decisão, regras de indução e visualização de dados.

A incerteza no gerenciamento da Rede é um fato, e a Mineração de Dados busca uma solução apoiada pela estatística Bayesiana para auxiliar na busca das tomadas de decisões no gerenciamento da rede, a qual interpreta a probabilidade como o

grau de certeza em um evento. As aplicações práticas de Redes Bayesianas (RB) vêm sendo bastante desenvolvidas no âmbito dos sistemas especialistas probabilísticos (principalmente na análise de comportamento aplicado à telefonia), da recuperação de informação probabilística, da descoberta de conhecimento em bases de dados (descrição), e da classificação bayesiana.

Neste contexto, mencionamos os Métodos Bayesianos que permitem representar quantitativamente o grau de certeza e manipular essas representações segundo as leis da probabilidade clássica. Com isso se obtém algumas vantagens na gerência de redes como:

- facilidade na administração, com a abrangência de todos os segmentos da rede;
- extração de conhecimento de uma amostra de dados proporcionando reações para os problemas obtendo assim uma possível prevenção para acontecimentos;

As habilidades desenvolvidas nas pessoas envolvem uma enorme quantidade de experiências do dia-a-dia e o aprendizado, acontece de forma bastante inconsciente. Nós aprendemos a falar nossa língua por intermédio da comunicação com os outros. Aprendemos a realizar tarefas por tentativas, às vezes com sucesso, às vezes falhando.

Em muitos casos, o conhecimento adquirido através dessas experiências não é consistente, mas, sem dúvida, se encontra em um formato de regras práticas, que armazenamos em algum lugar, para usá-las quando forem relevantes a uma nova situação. Essas regras práticas são chamadas de heurísticas, no linguajar de IA, e representam o suporte principal do conhecimento que tentamos armazenar para o uso dos sistemas em linguagem natural, sistemas especialistas e robôs.

O processo de obtenção de tais heurísticas a partir de uma pessoa e seu armazenamento em um computador é chamado de aquisição de conhecimento. Esta é indubitavelmente a parte mais difícil e que consome mais tempo, de qualquer projeto de Sistemas Baseados em Conhecimento.

A razão disso começa a ficar clara quando consideramos que o conhecimento de um ser humano inteligente é, na verdade, o somatório de toda a experiência de vida dessa pessoa.

Por conseguinte, para se ter uma máquina emulando segura e completamente uma determinada pessoa, seria necessário que ela obtivesse ‘toda’ a sua experiência de vida.

A partir de um determinado banco de casos, pode-se gerar um espaço de estruturas. Após encontrar todas as estruturas[1] com alta qualidade é necessário selecionar uma destas estruturas. Algoritmos de busca são empregados para este fim. Tais algoritmos, em geral, são “caros ” do ponto de vista de tempo de processamento.

A questão da construção de redes Bayesianas envolve dois aspectos: a construção da estrutura da rede e a definição das probabilidades associadas.

Na construção da estrutura, muitas redes são geradas e comparadas a fim de decidir qual é a melhor estrutura. Usam-se medidas de qualidade para comparar as estruturas: quanto “melhor” encaixar a estrutura com o banco de dados, maior a qualidade da mesma.

1.1. Objetivos

O objetivo principal deste trabalho é montar um Sistema Especialista Bayesiano que consiga detectar alterações de comportamento de dados em telefonia. Atualmente inúmeros celulares são clonados causando grande prejuízo para as empresas de telefonia. Para isso, serão simulados dados e realizados testes em ferramentas que executam clusterização de dados com o ARENA, NETICA e o STATISTICA. O programa *Bayeware Discoverer Student Edition* servirá como um dos modelos para a montagem da rede bayesiana.

Inicialmente o protótipo será montado em Delphi pela facilidade[2] e contato diário com a linguagem. A base dados simulada será modelada no MySql, um gerenciador de banco de dados com suporte a *scripts* SQL para a realização de busca de informações.

---

[1] o número de estruturas desses espaços cresce exponencialmente com o número de variáveis;

[2] programação orientada a eventos de componentes com suporte a Banco de Dados;

Os fundamentos de redes Bayesianas e seus métodos de propagação de evidências assim como aspéctos de inferência bayesiana, visarão tornar possível a construção de nós de conhecimento mapeando um padrão no uso de celulares e assim detectar, se houver, possíveis alterações no comportamento do uso dos celulares pelos clientes das operadores.

1.2. Objetivos Específicos

Desenvolver um protótipo para a análise de comportamento em telefonia celular. As ligações efetuadas por um cliente são registradas nas operadoras. Pela fragilidade na tecnologia celular a clonagem ocorre de maneira simples.

1.3. Estrutura do Trabalho

Este trabalho esta dividido em 4 capítulos. No primeiro faz-se a introdução, a apresentação dos objetivos assim como a estrutura do trabalho.

No capítulo 2 são apresentados conceitos de IA especificamente *Data Mining* (DM), fundamentos teóricos sobre probabilidade, redes bayesianas, onde é mostrada uma das possibilidades de extração de conhecimento através de inferência probabilística, reconhecimento de padrões assim como os trabalhos correlatos.

. No capítulo 3 trata da proposta do trabalho, das ferramentas adotadas para as simulações, clusterização e análise dos dados e construção de redes bayesianas. Também será abordado aspectos de aprendizado para a construção das redes.

No capítulo 4 são mostradas as contribuições do trabalho, o protótipo e o código fonte assim como as conclusões e melhorias com pesquisas em trabalhos futuros.

Capítulo II

Neste capítulo, serão descritas algumas etapas na produção de conhecimento em Banco de Dados (BD). Serão apresentados os conceitos de Mineração de dados como o processo de descoberta de conhecimento em BD.

As aplicações que trabalham com interpretação de informações precisam de métodos de análise de dados para que se possa toma determinadas decisões ou chegar a determinadas soluções. Em Redes Bayesianas usam-se cálculos que estimam estatísticas mostrando a probabilidade da ocorrência de um evento. Nas análises estatísticas são requeridos cálculos de freqüências.

2. Data Mining (DM)

No mundo dos negócios, eram usados para tomadas de decisões, os processos e métodos manuais para análise desses grandes volumes de dados através de planilhas eletrônicas e consultas diretas em *Structure Query Language* (SQL). Estes métodos de análise criavam relatórios informativos, mas não podiam analisar o conteúdo desses relatórios para enquadrar um conhecimento importante. Através desses fatos, surgiu a necessidade de novos métodos e ferramentas com habilidades para assistir os profissionais do mundo dos negócios nas análises dos grandes volumes de dados e para garimpar conhecimentos úteis.

Segundo Fayyad [FAY96], os avanços nas tecnologias de armazenamento de dados, tal como velocidade em acesso a dados, grande capacidade de armazenamento e barateamento de dispositivos de armazenamento, têm possibilitado gerar grandes volumes de dados. Com esses fatos, os Sistemas Gerenciadores de BD e tecnologias de *Data Storage* têm permitido transformar grandes quantidades de informações em uma montanha de dados armazenados.

Se faz necessário abordarmos conceitos de Inteligência Artificial (IA), especificamente *Data Mining* (DM), onde será mostrada uma das possibilidades de

extração de conhecimento através de inferência probabilística. Com um banco de dados de ligações de usuários, será possível o sistema detectar, , através de um monitoramento em uma base de dados, os perfis de usuários e principalmente as possíveis modificações de comportamento através da rede *a priori*.

Com o rápido progresso nas pesquisas de DM e *Data Warehouse* (DW), nos últimos cinco anos, muitos sistemas de DW e DM tem sido desenvolvidos na extração de conhecimento de banco de dados relacionais; um importante problema na mineração de dados consiste em descobrir uma associação de regras de um bancos de dados de transações onde cada transação consiste em um conjunto de itens. Muitas vezes realizamos operações na descoberta de processos computacionais de ocorrências de freqüências em um sub-conjunto de itens de um banco de dados de transações.

O usuário inicia o processo de descoberta de conhecimento que apresenta características interativas e iterativas, envolvendo muitos passos com muitas decisões. A extração de conhecimento é um processo que limita sua atenção para os dados que deverão ser guardados e visualizados das instâncias que serão extraídas da complexidade do mundo real.

A mineração de dados é uma etapa crucial dentro dos muitos passos do processo aquisição do conhecimento. A utilização da mineração de dados traz algumas vantagens para encontrar padrões estatisticamente significantes mesmo em dados com características casuais. O sistema de *Knowledge Discovery Data* (KDD) pode apresentar muitos problemas na mineração de dados e deve proteger os usuários desses perigos. Entretanto, mesmo com esses riscos, pode-se ter uma esperança de retorno válido na mineração.

Segundo Fayyad [FAY98], na mineração de dados, são vistos interessantes estruturas entre os dados, tornando a forma de padrões ou modelos. O ato de encontrar padrões em dados tem sido estudado ao longo do tempo em outros campos – como estatística, reconhecimento de padrões e dados.

Existem muitas tecnologias e ferramentas disponíveis para aplicações em DM. Certas tecnologias tem melhor desempenho que outras em termos de facilidade de

uso e retorno do investimento, porém na maioria das vezes uma ferramenta sozinha nunca apresentará uma solução completa.

2.1. Processo de Extração de Conhecimento de BD

A idéia geral é que fora as situações em que se possa empregar métodos matemáticos padrão ou análise estatística para testar hipóteses pré-definidas, DM é a solução mais útil na análise exploratória em cenários onde não existem noções predeterminadas sobre o que irá constituir um saída interessante. *Data mining* é um processo iterativo onde o progresso é definido pela descoberta, quer seja automática ou manual [WES98].

Abaixo descreve-se os passos básicos do processo de extração de conhecimento segundo Fayyad [FAY96]:

- Desenvolvimento da compreenção do domínio da aplicação: ressalta o conhecimento prévio dos objetivos de um usuário final.
- Criando objetivos para os conjuntos de dados: seleciona-se o conjunto de dados, ou concentra-se em um subconjunto de variáveis ou exemplos de dados, nos quais a descoberta será aplicada. O processo KDD inicia-se com a seleção dos dados alvos, o subconjunto de armazéns de dados relevantes para a tarefa. A seleção é geralmente acompanhada por uma limpeza de dados [FAY98].
- A limpeza dos dados e pré-processamento: esta operação basicamente retira os ruídos ou sujeiras, coletando as informações necessárias para a modelagem por causa dos ruídos, decidindo as estratégias para tratamento de ausências de dados, informações e conhecimentos modificados. Segundo Fayyad [FAY98], enquanto algumas limpezas podem ter sido feitas no estágio de armazenamento, este passo envolve operações relatadas explicitamente para a tarefa de mineração de dados, decidindo como tratar perdas de valores de campos, sumarizando múltiplas linhas dentro de entradas simples, detectando casos distorcidos. Uma importante parte desta tarefa é como identificar diferentes transações relatadas para algumas entidades. Cada identificação pode ser feita por comparação de registros

similares e levantando-se todas as informações relevantes sobre algumas entidades para facilitar os próximos passos.

- Redução de dados e projeção: este passo apresenta características proveitosas para representar os dados, dependendo dos objetivos das tarefas. Usando reduções dimencionais ou métodos de transformações para reduzir o número efetivo de variáveis ou para encontrar constantes representativas para os dados. Segundo Fayaad [FAY98], este passo envolve transformação do dado dentro de formas apropriadas para o método de mineração de dados. Transformações podem incluir reduções de passos de dimensionalidade (substituindo alguns campos com um campo derivado) ou passo de redução de dado (mapeando múltiplos registros para uma simples entidade e derivando resultados dentro de conjuntos de registros). Também pode adicionar novos registros que o usuário pode julgar relevantes para o problema. O resultado é um grande conjunto de dados e tipicamente envolve unificação de dados dentro de relacionamento simples de banco de dados.

- Escolhendo as tarefas de mineração de dados: considerando os objetivos do processo de aquisição de conhecimento, decide-se qual das tarefas será utilizada como por exemplo, classificação, regressão, clusterização, etc.

- Escolhendo os algoritmos de mineração de dados: seleciona-se o método para ser usado pelas buscas de padrões de dados. Este passo decide qual modelo e parâmetro que podem ser apropriados e compatíveis em um particular método de mineração de dados para um completo critério de um processo de aquisição de conhecimento.

- Mineração de dados: buscas padrões de interesse em um particular meio representacional ou um conjunto de cada representação, através da utilização de classificações de regras de associação, regressão, clusterização, classificação e assim por diante. O usuário pode ajudar significativamente os métodos de mineração de dados, através de correções no desempenho nos passos precedentes. Na fase de mineração de dados, um algoritmo é aplicado para encontrar estrutura de interesse (padrões ou modelos) nas tabelas de dados. O padrão poderia ser um modelo previsível, sumarização de dados, segmentação

de dados ou um modelo de dependência (ligações) dos dados. O tipo de padrão gerado depende das tarefas de mineração de dados. Para um conjunto finito de dados de grande volume, números possíveis de padrões e modelos podem ser extraídos. O melhor método de busca fica impraticável sem as técnicas de busca heurística[3].

- Interpretando as extrações de padrões: possibilita retornos para alguns dos passos anteriores para promover iterações. Segundo Fayaad [FAY98], este passo avalia a seleção de padrões para os objetivos, decidindo quais deles são interessantes para interpretá-los legivelmente para o usuário ou tornar aceitáveis para as aplicações que usarão os dados. O critério de avaliação inclui: validação, utilização, descoberta, entendimento (simplificado) e grau de interesse.
- Consolidação da descoberta de conhecimento: incorporando estes conhecimentos dentro das execuções dos sistemas, ou simplificando documentações desses conhecimentos e reportando-os para interesses particulares. Também inclui checagem para resolver conflitos potenciais de valores previamente conhecidos.

A extração do conhecimento pode envolver significantes iterações e pode conter loops entre qualquer um dos passos sendo que a descoberta de conhecimento será usada por pessoas com o intuito de aumentar seus entendimentos ou integrá-los com os processos, para desempenhar previsões a fim de tomada de decisão ou controle. Este é um passo crucial e muito desafiante [FAY98].

2.2. Métodos de Mineração de Dados

Os componentes da mineração de dados do processo descoberta de conhecimento muitas vezes são envolvidos por repetidas iterações aplicadas por um particular método de mineração de dados [FAY96].

A mineração de dados utiliza combinações para um determinado padrão de observação de dados. O modelo combinado deixa a cargo do conhecimento por

[3] conjunto de métodos que auxiliam na resolução de problemas

inferência: se o modelo reflete ou não o conhecimento útil ou interessante, o processo extração de conhecimento interativo requer o julgamento humano de forma subjetiva. Existem dois formalismos usados em um modelo combinado: o estatístico somente para efeitos não determinísticos e o modelo lógico que é puramente determinístico e não admite a possibilidade de incertezas na modelagem do processo.

As atividades de mineração de dados são descritas dentro de cinco categorias [FAY98]:

- Previsão de modelos: O objetivo é prever alguma situação de algum campo alvo dos dados, usando os restantes dos campos. Quando a variação inicial prevista é categórica (para aprovar/rejeitar um empréstimo, por exemplo), o problema é chamado classificação. Quando a variável é contínua (tal como a forma esperada de ganho/perda), o problema referido é uma regressão.
- Classificação: Simplesmente acessa técnicas de padrões de inclusões para regressão linear e análise discriminante. Métodos populares dessa categoria de mineração de dados inclui árvores de decisões, regras, redes neurais (regressões não lineares), funções básicas radiais, e muitos outros.
- Banco de dados clusterizados (segmentação): são encontrados os clusters nos dados que consistem de subconjuntos de registros similares. Diferentes dos modelos preventivos, não existe nenhuma variável alvo. Existem muitos métodos para clusterização de dados. O objetivo da clusterização é tomar um conjunto de entidades representativas como registros em banco de dados e particioná-lo dentro de um número de grupos ou clusters, aonde cada cluster contém registros similares.
- Dependência de modelos: o objetivo é entender os resultados probabilísticos das densidades dos dados. Se é conhecida a função densidade, pode-se responder algumas questões de interesse sobre os dados. Estimar densidade é um problema difícil e pode ser extremamente complicado em grandes dimensões de conjunto de dados. É possível se obter um método escalar com clusterização e densidade estimada.

- Sumarização de dados: envolve métodos para encontrar uma compacta descrição para um subconjunto de dados.

2.3. Abordagem para Mineração de Dados

Segundo Menezes [MEN98], uma variedade de métodos e técnicas usadas correntemente para extração de conhecimento de grandes volumes de dados, vem de diversas áreas como estatísticas, aprendizagem de máquinas, banco de dados, visualização, e representação de conhecimentos para sistemas inteligentes. A abordagem para mineração de dados podem ser classificadas nas seguintes categorias:

- Abordagem estatística: Nas décadas passadas, a estatística tem desenvolveu métodos usados para avaliar hipóteses e determinar as diferenças que podem ser relacionadas a chances aleatórias. A teoria da estatística formal suporta modelos de dados e métodos de previsão. Neste contexto, inferência bayesiana é o mais usado método estatístico para descoberta de conhecimento. Três métodos bayesianos usados para problemas de mineração de dados são Naive Bayes Classifier (Mitchell), Autoclass (Cheeseman & Stutz), e Bayesian Networks (Heckerman).
- Abordagem de aprendizagem de máquina: Abaixo desta categoria, são incluídos métodos de predição desenvolvidos por comunidades da área da ciência da computação, para especificar alguns modelos de interesse. Alguns dos algoritmos mais comuns de aprendizagem de máquinas usados para minerar conhecimento de conjunto de dados, são K-Nearest Neighbour (Dasarathy), Decision Trees (Quinlan), Neural Networks (Hertz) e Genetic Algoritms (Golberg).
- Abordagem para banco de dados orientado: Banco de dados é o repositório com numerosas transações que envolvem os negócios comerciais e a comunidade científica. O desenvolvimento de armazéns de dados e ferramenta *Online Analytical Processing* (OLAP) para a análise de grandes BD, têm encorajado o aparecimento de outros métodos para mineração de conhecimento de diferentes

tipos de banco de dados. Algumas dessas técnicas são Attribute-Oriented Induction (Han & FU) e Association Rules (Agrawal).

- Abordagem de visualização: As técnicas de visualizações estão se tornando métodos muito usados para descoberta de padrões em conjunto de dados, porque eles impactam diretamente nos sistemas visuais humanos, correntemente o mais forte identificador de descoberta de padrão. Existe uma ampla variedade de técnicas, que podem ser usadas em alguns estágios do processo de extração de conhecimento: no estágio de pré-processamento, para conseguir amplo sentido nas características dos conjuntos de dados; no estágio da mineração de dados, para descobrimento de padrões; para visualizar o modelo produzido por algoritmo de mineração de dados, com a finalidade de melhor entender as responsabilidades geradas por um modelo. Dentro das técnicas mais comuns de visualização para mineração de conhecimento de dados, destacam-se: 2D e 3D , Multidimensional visualization techniques, hierarchical techniques.

- Abordagem de representação de conhecimento: É a representação de extração de conhecimento por um sistema de mineração de dados, envolvendo uma troca de forças expressivas e complexidades computacionais. Dentro das mais comuns representações de conhecimento em sistema de mineração de dados, estão: Propositional-like representations; produção de regras (Quinlan); lista de decisão(Rivest); primeira ordem lógica (FOL) e o FOIL (Quinlan) são exemplos de sistema de programação lógica indutiva que usa FOL para representar conhecimento.

2.4. A descoberta de conhecimento em aplicações de negócios

Segundo Brachman [BRA96], em termos tecnológicos, um sistema de descoberta de conhecimento, poderia ser um dispositivo integrado que de certo modo assiste o usuário na condução para o processo de descoberta de conhecimento. As noções comuns de saída de um sistema de descoberta de conhecimento são simplesmente alguns fragmentos de conhecimento e a saída de um processo de descoberta de conhecimento em um sistema comercial, no mínimo, deveria ser uma especificação de uma aplicação de descoberta de conhecimento. Cada uma das

aplicações poderão ser construídas e instaladas em um dispositivo comercial, para disponibilizar análises e ações recomendadas em uma base avançada, usando como exemplo, dados de negócios. Estes usuários deverão ser pessoas de negócios dando atenção para importantes eventos em dados de negócios e, particularmente, um analista de dados explorando a base geral e padrões em um domínio. Este domínio pode ser entendido como classes de empresas de negócios, áreas científicas, governamentais e pessoas de negócios comerciais.

Os processos de extração de conhecimento em aplicações de negócios, dividem-se em três tarefas principais: os dois primeiros itens são âncoras gerais para analisar e visualizar métodos suportados por ferramentas de consultas de banco de dados. O terceiro item é geralmente suportado por uma variedade de apresentações e ferramentas de transformação de dados.

2.4.1. Análise de dados

Este processo é visto como análise confirmatória dos dados, é subdividido em:

- Especificação de modelos: no qual um especificado modelo é escrito fora de um caminho normal. Um modelo inicial pode ser derivado de muitos modelos diferentes.
- Combinação de modelos: quando necessário, algumas especificações de parâmetros de modelos são determinados e baseados nos dados (em alguns casos, o modelo é construído independentemente dos dados, enquanto em outros são combinados para instruções de dados).
- Avaliação: onde os modelos avaliam os dados (em alguns casos, todos os dados; em outros, somente os dados testados).
- Refinamento de modelos: onde o modelo inicial pode ser iterativamente refinado, dependendo do que mostra a avaliação. Existem alguns tipos de ferramentas que suportam esses subprocessos. Essas ferramentas se caracterizam em análise com base algoritmica e base de visualização. Base algoritmica é um modelo que é especificado por meio de associação de um resultado variável . Em certos casos, o conjunto de dados é subdividido dentro de um teste de conjunto

de dados. A preparação conjunta de dados é então usada para combinar os parâmetros de um modelo. Este modelo é então avaliado e aplicado para um conjunto de dados, e se necessário, sofre subseqüentes refinamentos. Estas ferramentas de bases algoritmicas incluem certas ferramentas de aprendizagem de máquinas (árvores de decisão indutiva e outras supervisões de técnicas de aprendizagem). Redes neurais são um caso baseado em plataformas de raciocínios e ferramentas de classificação. Também as ferramentas de inclusões de pacotes estatísticos. Ferramentas de análise com base de visualização e hipótese são específicas por meio de uma visualização comparativa (por exemplo um tipo específico de gráfico) e a seleção de um dado elemento para ser usado na comparação. A visualização produzida é por si mesma um modelo. Em uma tarefa de descoberta de conhecimento realístico, a visualização é um ingrediente chave para todas as três fases centrais do processo (desenvolvimento de modelo, análise de dados, e gerações de saída). Para algumas tarefas, visualização apropriada é somente necessária para esclarecer um problema ou confirmar uma hipótese, mesmo que geralmente não se pense na figura traçada como uma qualidade de operação.

2.4.2. Desenvolvimento de Modelos

Raramente, o analista inicia uma hipótese de especificação formal para ser confirmada ou não confirmada. Em muitas aplicações de descoberta (por exemplo, análise de dados de mercado), a população total pode ser complexa para a compreensão, mas detalhes sobre os subconjuntos de uma população são mais manejáveis. Similarmente, mesmo quando se conhece a sub população que se quer analisar, pode ser necessário restringir parâmetros usados para fazer a análise – nem todas as variáveis serão de utilidade nesta análise, estas variáveis poderão ser uma relação correlata entre elas que serão necessárias para correção e a completa quantia de dados pode ser também decidida para tratamento.

Todos esses significados que existem são importantes no processo de descoberta de conhecimento que podem anteceder a atual análise de dados.

Esta parte da extração de conhecimento é dividida em subprocessos como:

- Segmentação de dados: Dentro das ajudas mais comuns para segmentação de dados são técnicas de aprendizagem (clusterização).
- Seleção de modelo: Existe uma ampla variedade de modelo de análise que poderá ser usado em grandes quantidades de dados, incluindo regressão, árvores de decisão, redes neurais, e raciocínio baseado em casos. O analista tem que escolher o melhor tipo de modelo antes de invocar alguma ferramenta particular de análise; geralmente ele explora diferentes tipos de modelos antes de decidir aonde mostrar o melhor para a utilização.
- Seleção de parâmetro: Finalmente, dentro do modelo escolhido, deve ser determinado algum parâmetro necessário para enfoques sucessivos.

Capítulo III

As empresas possuem Bases de Dados com informações que podem vir a se repetir possibilitando a geração de probabilidades das futuras ocorrências destas informações. O conhecimento à priori pode ser convertido em algoritmo com previsões probabilísticas de hipóteses daquelas ocorrências. A Bayes deve-se o cálculo das chamadas probabilidades e das causas. Ou seja, este cálculo consistiu em determinar a probabilidade de acontecimentos perante certas condições iniciais.

3. Teorema de Bayes

Redes Bayesianas (RB) são mecanismos eficientes para análise de dados que apresentam relacionamentos de precedência temporal. Uma RB tem dois componentes: uma estrutura gráfica e os parâmetros numéricos. Tanto a estrutura gráfica quanto os parâmetros numéricos podem ser aprendidos automaticamente de uma base de dados. Este trabalho dá a base teórica e propõe um conjunto de algoritmos e estrutura de dados que permitem manipular grandes volumes de dados no processo de aprendizagem de uma RB. Enfoca a limitação teórica dos processos de aprendizagem e detalha o principal algoritmo de aprendizagem Bayesiana que explora o método de independência condicional.

Em 1762, o Reverendo Thomas Bayes demonstrou um procedimento bastante importante para se calcular a probabilidade de um evento dado que um outro tenha ocorrido. O teorema de Bayes tem importância fundamental, pois fornece a base para uma abordagem da inferência estatística. O seu método consiste na partição do espaço amostral em diversos subconjuntos cujas probabilidades sejam conhecidas e, em seguida, aplicar a seguinte fórmula:

$$P(Ai/B) = \frac{P(B/Ai).P(Ai)}{\sum \mathrm{P(B/Aj).P(Aj)}} \qquad (1)$$

Na equação acima, *Ai* é o evento cuja probabilidade se quer calcular dado que o evento B tenha ocorrido. *Aj* representa genericamente uma das "fatias" da partição do espaço amostral (*Ai* é uma dessas fatias).

A interpretação da probabilidade como o grau de certeza em um determinado evento, é chamado “enfoque Bayesiano”. Neste enfoque, quando não se pode mensurar através de dados ou fatos a freqüência de um evento qualquer, Sua probabilidade pode ser dada com base no julgamento humano desde que se possa garantir que este julgamento soja feito de forma honesta [Pearl88]

A famosa Regra de Bayes, como é conhecido o método, permite que se ajuste uma probabilidade *a priori* (conhecida) de um dado evento à luz de novas evidências envolvendo um outro evento que apresenta relação de dependência com o primeiro.

Os métodos Bayesianos possibilitam representar numericamente o grau de certeza sobre um evento, e manipulação de acordo com as regras definidas na teoria de probabilidade. Pois a teoria Bayesiana está fundamentada na teoria da probabilidade, sendo que a diferença básica está no enfoque não freqüencista adotado pela teoria de Bayes. Na teoria da probabilidade, dados dois eventos A e B, é possível condicionar A a ocorrência de B. Essa probabilidade condicional é dada por:

$$P(A|B) = P(A \cap B) / P(B), \quad \text{se } P(B) > 0 \qquad (2)$$

P(A|B) pode ser interpretada como a atualização da crença em A dado que a evidencia B tornou-se disponível. Em outras palavras, P(A|E) representa a probabilidade do evento *A* (uma hipótese) condicionado à ocorrência de algum evento B (evidencia). Se a intenção é encontrar o número de pontos do espaço amostral, na verdade busca-se a fração de eventos B que também estejam contidos em A. O teorema de Bayes descrito pode facilmente ser reescrito de forma a obter a probabilidade posterior de uma hipótese A, após a observação de alguma evidencia A, dado a probabilidade *a priori* em *A* e a verossimilhança da observação B estar em A:

$$P(A|B) = (P(B|A)\ P(A)) / P(B); \quad \text{se } P(B) > 0 \qquad (3)$$

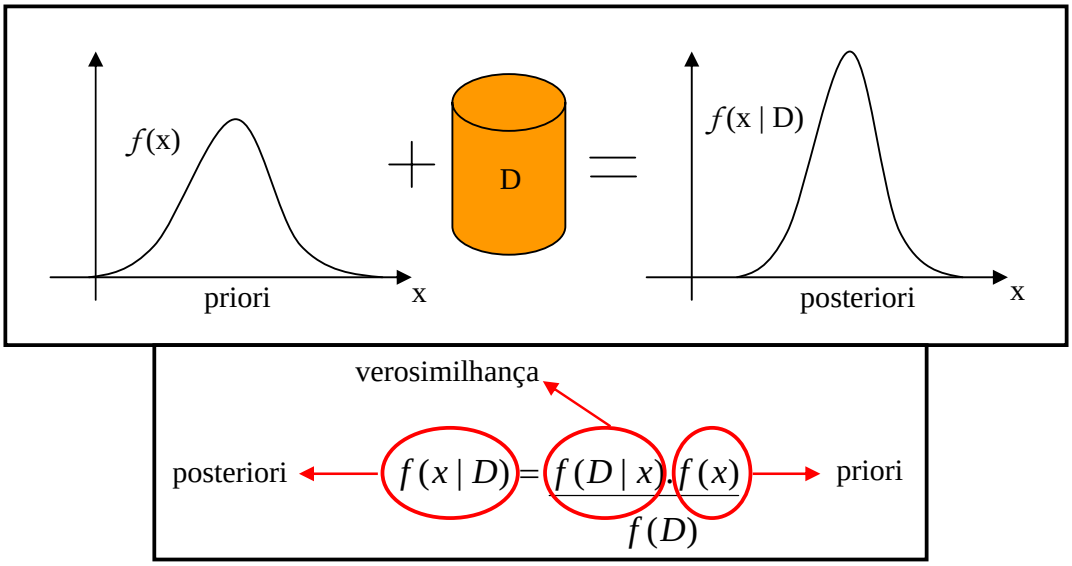
Esta fórmula simples tem imensa importância prática num domínio como o de diagnóstico. É freqüentemente mais fácil deduzir a probabilidade, por exemplo,

observando um sintoma dado a doença do que a doença dada o sintoma. O teorema de Bayes pode ser usado para manipular o caso onde a hipótese é uma proposição no domínio do conhecimento (uma doença específica, talvez) e a evidencia e a observação de algumas condições (talvez um sintoma). Todavia, o teorema pode também manipular casos onde urna hipótese é um parâmetro numa rede tendo um certo valor (ou distribuição de valores) ou que a rede tenha uma certa estrutura, e a evidencia dados que chegam sobre o caso.

3.1. Inferência Bayesiana

É um método que nos permite calcular a densidade *posteriori* de uma variável aleatória desejada em base a sua densidade *priori* e o conjunto de dados.

Figura 1: Análise de Probabilidades



A figura 1 mostra como o protótipo vai atuar nos dados a partir de estatíticas *a priori* e o estabelecimento do vínculo entre os valores observados em uma amostra e os modelos postulados para uma população, objeto da inferência bayesiana. Com os dados dos clientes é possível o estabelecimento de probabilidades para futuras decisões, com isso uma rede bayesiana pode ser estabelecida através das probabilidades *a priori* e

com as aproximações dos futuros dados, projetar as probabilidades *a posteriori* para possibilitar o fluxo pelos nós (probabilidades) da rede bayesiana formada.

3.2. Redes Bayesianas

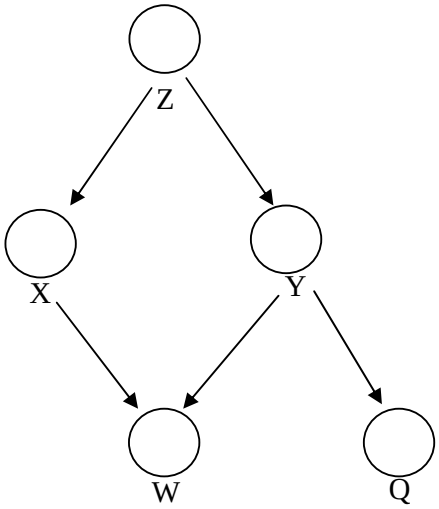
Uma rede bayesiana, por exemplo, é um grafo acíclico orientado, onde os nós representam variáveis aleatórias e o arco unindo dois nós representam a dependência probabilística entre as variáveis associadas. Cada nó possui armazenada a função de distribuição de probabilidades condicional dos valores que podem ser assumidos pela variável aleatória associada ao nó, dado os valores de seus nós pais (isto é, aqueles diretamente ligados ao nó em questão).

As Redes Bayesianas são tipos específicos de redes de conhecimento. A idéia principal é que, para descrever o mundo real não é necessário usar uma enorme tabela de probabilidades conjuntas na qual listamos as probabilidades de todas as combinações possíveis de eventos. A maioria dos eventos é condicionalmente independente da maioria dos outros, portanto suas interações não precisam ser consideradas. Em vez disso, podemos usar uma representação mais local em agrupamentos de eventos que interagem. O primeiro pesquisador a idealizar a utilização da representação gráfica das informações probabilísticas foi o geneticista Scwallwrigth, mas foi muito criticado e o seu trabalho se foi reconhecido pela comunidade estatística a partir da década de 50. Estas redes foram desenvolvidas nos anos 70 com objetivo de modelar processamento distribuído na compreensão da leitura, onde as expectativas semânticas e evidências perceptivas deveriam ser combinadas para formar urna interpretação coerente.

Uma rede Bayesiana é um grafo direcionado acíclico onde os nós representam as variáveis (de interesse) de um domínio e os arcos representam a dependência condicional ou informativa entre as variáveis. A força da dependência é representada por probabilidades condicionais que são associadas a cada grupo de nós pais-filhos na rede. Para verificar se uma rede é uma rede Bayesiana, basta que uma condição seja satisfeita: cada variável do grafo deve ser condicionalmente independentes de todos os nós que não são *sets* descendentes exceto seus pais. Dada a distribuição de probabilidade conjunta $P(x_1, x_2, x_3, ..., x_n)$ e uma determinada

ordenação *d* destas variáveis, inicia-se a construção do grafo, escolhendo o nó raiz X_1 e atribuindo a ele a probabilidade marginal *P(x,)*. Em seguida, acrescenta-se mais um nó X_2 no grafo. Se X_2 for dependente de X, então traçamos um arco ligando ambas as variáveis com a seta apontando para X_2 e quantifica-se este arco com $P(x_2|x_1)$, caso contrário, mantém-se as variáveis desconectadas e atribui-se uma probabilidade à priori $P(x_2)$ a X_2. Repete-se a operação para as demais variáveis e, então, obtém-se uma rede Bayesiana.

Figura 2: Rede Bayesiana



A rede acima representa a seguinte distribuição:

$$P(Z,X,Y,W,Q) = P(Q|Y)\ P(W|\ KV)\ P(Y|\ Z)\ P(X|\ Z)\ P(Z)$$

As relações de independência podem reduzir o esforço no cálculo da distribuição de probabilidade conjunta. Mas nem sempre é possível a identificação das relações de dependência entre as variáveis. Nesses casos usa-se o julgamento humano para defini-las. Nos casos citados, deve-se buscar a relação de causalidade entre as variáveis e, para aquelas que influenciam diretamente outras, definir pai e filho, respectivamente.

Propagar evidência em redes Bayesianas corresponde a estabelecer um procedimento que, explorando a conectividade da rede, permita a determinação das distribuições de probabilidades das variáveis objetiva do diagnóstico e/ou prognóstico,

condicionadas aos valores das variáveis que reapresentam, evidências observadas. A propagação é um mecanismo que permite avaliar as probabilidades posteriores das variáveis da rede. Para se manter a consistência da crença sobre as variáveis do modelo, sempre que surge uma nova evidência a favor de uma variável a crença nesta variável e alterada e esta nova crença propagada por toda a rede. Com isso, toda a rede atualiza sua crença e os efeitos da nova evidência são absorvidos por todas as variáveis, fazendo com que a representação continue consistente. [HRU98]

O teorema de Bayes é um método quantitativo para a revisão de probabilidades conhecidas, com base em nova informação amostral. A grande diferença entre o método Bayesiano e a probabilidade clássica é a inclusão de um elemento de subjetividade que o primeiro oferece.

A estatística Bayesiana é a base das redes probabilísticas causais e um ferramental poderoso já validado.

As Redes Bayesianas são redes de conhecimento, representadas por grafos direcionados acíclicos, onde os nós representam variáveis aleatórias com medidas de incerteza associadas e os arcos representam a interdependência entre estas variáveis, e possuem uma quantificação da força deste relacionamento. Nessas redes, pode-se calcular a probabilidade de um evento ocorrer condicionado à ocorrência de outro. Elas representam a incerteza tomando por base a teoria da probabilidade. Com o aparecimento de modelos de redes probabilísticas, a teoria da probabilidade surgiu novamente com grande vigor na IA para modelar incerteza. [HRU98]

A questão de construção de redes Bayesianas envolve dois aspectos : a construção da estrutura da rede e a definição das probabilidades associadas.

Construir redes de conhecimento Bayesianas com auxílio de especialistas é um problema que envolve grande consumo de tempo e, em muitos casos, o próprio domínio não é bem conhecido. Além do mais, especialistas humanos são bons em julgar relações qualitativas entre as variáveis de um domínio, mas não tão bons em quantificar as probabilidades associadas. Daí, processos de geração automática de redes serem implementados para auxiliar este processo de aquisição do conhecimento pela exploração, entre outras, das informações constantes em bancos de casos.

Assim, as redes encontradas por esses algoritmos tornam-se ponto de partida para a fase de teste da representação do conhecimento acerca de um determinado domínio, o que facilita e diminui consideravelmente o tempo de construção da rede.

Na construção da estrutura, muitas redes são geradas e comparadas a fim de decidir qual é a melhor estrutura. Usam-se medidas de qualidade para comparar as estruturas: quanto “melhor” encaixar a estrutura com o banco de dados, maior a qualidade da mesma.

A partir de um determinado banco de casos , pode-se gerar um espaço de estruturas. Após encontrar todas as estruturas com alta qualidade – o número de estruturas desses espaços cresce exponencialmente com o número de variáveis – é necessário selecionar uma destas estruturas. Algoritmos de busca são empregados para este fim. Tais algoritmos, em geral, são “caros ” do ponto de vista de tempo de processamento. Métodos heurísticos são amplamente utilizados, visando minorar o tempo de execução.

É importante ressaltar que uma rede completa ainda necessita de probabilidades associadas à estrutura selecionada. Estas probabilidades podem ser estimadas diretamente a partir do banco de dados.

A estatística Bayesiana tem origem no trabalho de Thomas Bayes, mas na verdade foi o matemático francês Pierre Simon de LaPlace (1812) quem desenvolveu o teorema na forma como ele é conhecido e utilizado atualmente .

3.3. Dependência e Independência de Eventos

Quando se estuda simultaneamente dois eventos, existem duas possibilidades quanto à relação entre as suas probabilidades:

- Eventos Dependentes: quando a ocorrência de um evento influencia a probabilidade de ocorrência do outro;
- Eventos Independentes: quando a ocorrência de um evento em nada interfere na ocorrência do outro.

Assim, se A e B são eventos independentes, então a probabilidade de que ambos aconteçam ao mesmo tempo é necessariamente igual à probabilidade isolada de

um deles ocorrer multiplicada pela probabilidade isolada do outro, ou seja, em notação matemática:

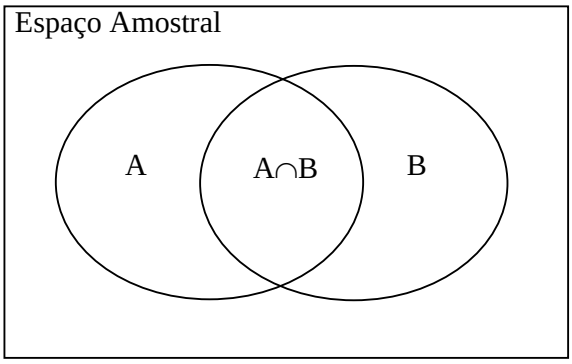
$$P(A \cap B) = P(A) \times P(B)$$

Quando existe alguma relação entre A e B de modo que a ocorrência de um interfere na probabilidade do outro, a probabilidade de ambos ocorrerem ao mesmo tempo assume um valor diferente dependendo da natureza da relação.

3.4. Probabilidades de Eventos Depedentes

Quando se tem dois eventos que se mostram dependentes, o cálculo das chances de cada um envolve as chamadas Probabilidades Condicionais.

Figura 3: Espaço Amostral



A probabilidade de que ocorrências dentro do espaço amostral correspondam ao evento A é ilustrada na figura 3 pela área da elipse interna. A probabilidade no caso do evento B é dada pela área da elipse contínua. A probabilidade de que as ocorrências venham a corresponder tanto a A quanto a B é dada pela área da interseção entre as duas elipses (A∩B). Assim, a probabilidade de A dado que B ocorreu é calculada pela razão entre a interseção entre A e B e a área de B, ou seja, em notação matemática,

$$P(A|B) = P(A \cap B)/P(B). \qquad (1)$$

Naturalmente, é possível se usar o mesmo raciocínio para se calcular a probabilidade de B dado que A tenha ocorrido, isto é,

$$P(B|A) = P(A \cap B)/P(A). \quad (2)$$

3.5. Independência Condicional

Diz que duas variáveis A e B são independentes se

P(a/b)=P(a), sempre que P(b)>0, ∀a∈Da e b∈Db

Se A e B são independentes, então B não é informativo para X. Significa que conhecer B, não altera a probabilidade de A. Pode-se expressar essa independência em termos da distribuição conjunta de A e B, derivada da regra do produto, equação (2), como

$$I(A,B) = \sum_{A}\sum_{B} P(a,b)\log\frac{P(a,b)}{P(a)P(b)} \quad (3)$$

Uma outra maneira de fazer esse teste de independência é usar a medida de informação mútua, dada pela equação (3). Essa relação é reflexiva, I(A,B) = I(B,A). Se I(B,A) > 0, então A e B são informativos um para o outro, isto é, são dependentes; em

$$I(A,B,C) = \sum_{A}\sum_{B}\sum_{C} P(a,b,c)\log\frac{P(a,b/c)}{P(a/c)P(b/c)} \quad (4)$$

caso contrário, A e B são independentes. É fácil ver que se A e B são independentes, o argumento da função log, na equação (4), é igual a 1, para todos os valores de A e B, o que leva o somatório ao valor 0(zero).

Dado um subconjunto de variáveis aleatórias C, não contendo A nem B, pode-se também usar a medida de informação mútua condicional para verificar se A e B são condicionalmente independentes, conforme a equação (4). Como no caso anterior, se I(A,B|C)=I(B,A|C)>0, então A e B são condicionalmente dependentes, e são condicionalmente independentes em caso contrário. Em termos probabilísticos, se P(a/b,c)=P(a/b), então conhecido C, conhecer B não informa nada para A. Isto é, A é independe de B, dado C. Também é verdade que P(a,b/c) = P(a/c)P(b/c), se A e B são independentes, dado C.

3.6. Trabalhos Correlatos

Um sistema especialista é perfeitamente associado na área de gerência pró-ativa, podendo ser complementado com outras técnicas de inteligência artificial. Diversas técnicas de inteligência artificial estão sendo pesquisadas e aplicadas na área de gerência pró-ativa. Como em [FRA97], que através dos dados colhidos por monitoração remota desenvolveu um sistema de gerência pró-ativa com uso de inteligência artificial.

Em [CAR00], dois algoritmos mais significativos ao processo de aprendizagem são comparados , após aplicação a uma base de casos comum. Um algoritmo pertencente à classe dos métodos de busca e pontuação utilizando um Método de Pontuação Bayesiana[4] – especificamente o algoritmo K2 (ANEXO 1) – é implementado e aplicado a uma base de casos.

Em [HRU98], modela-se o conhecimento do domínio de doenças pulmonares; propagar evidências sobre o modelo obtido e avaliar a qualidade da inferência probabilística realizada sobre o modelo. Abordam-se métodos de propagação de evidências com suas características e seus algoritmos detalhando-se o comportamento de dois algoritmos refinando o resultado com auxílio de um especialista.

[4] os algoritmos vêem o problema “aprender” como um problema de busca pela estrutura que melhor se encaixa com os dados, iniciando com um grafo sem arcos utilizando algum método de busca para adicionar um arco ao grafo para posterior utilização de um método de pon tuação para identificar se a nova estrutura é melhor do que a antiga.

Capítulo IV

Você começa a receber muitos telefonemas por engano e o aparelho não completa as ligações com a mesma facilidade de antes. Esses são sinais de que o celular pode ter sido clonado. "O consumidor tem direito ao ressarcimento das ligações não reconhecidas e pode pedir o cancelamento do serviço, com a devolução da mensalidade", afirma Gabriela Ribas Antônio, assessora de diretoria do Procon. É necessário que o assinante prejudicado faça um boletim de ocorrência para se proteger de eventuais disputas com a operadora.

Segundo Guilherme Guimarães, em entrevista da Editora Abril, o celular tornou-se tão popular no nosso dia-a-dia que já alcançou uma posição de destaque na lista de reclamações do Procon, aparecendo em terceiro lugar, atrás de operadoras de cartões de crédito e planos de saúde. Umas das principais queixas diz respeito à clonagem do número dos aparelhos.

A necessidade de identificação de um perfil de usuário se faz necessário pelo auto número de fraudes ocorridas nas prestadoras de serviço. Um exemplo são as fraudes de assinatura, na qual o usuário faz inúmeras ligações mas com o contrato da linha em nome de outra pessoa mediante a suborno e fraudes por clonagem do aparelho celular. Clonar um celular significa copiar a freqüência de um aparelho para um outro, permitindo ao criminoso utilizar seu número sem autorização. Esta fraude ocorre quando o celular opera em "roaming" (fora da área de origem do aparelho) e em modo analógico. Quando o aparelho é ligado, ele varre a área de cobertura para buscar a melhor banda de atuação. Se não houver cobertura digital, o celular funciona no modo analógico, ou seja, transmite os dados por ondas de rádio, capturáveis pelos criminosos e configuradas em outro aparelho, o clone.

4. Domínio da Aplicação

Este trabalho explora a aplicação de um Sistema Especialista Bayesiano no apoio à tomada de decisões, particularmente na detecção do perfil de usuários de telefones e detecção de auterações mediante modifcação dos dados no BD.

4.1. Protótipo

A heterogeneidade de informações em uma base de dados pode ser uma dificuldade para o sistema, no que se refere a encontrar a melhor distribuição matemática para análise, obtenção de uma "atitude" como detectar um perfil de usuário. A base de dados utilizada inicialmente foi um arquivo de ligações da TIM onde se encontram algumas métricas como a duração[5], destino[6]. Outros critérios como desconto e dia da semana são interessantes, de acordo com esclarecimento de um funcionário da empresa, para a formação do perfil do usuário.

4.2. Banco de Dados

O banco de dados MySQL é um banco de dados relacional gratuito, eficiente e otimizado para aplicações Web. Por esta razão, ele possui menos recursos, se comparado com outros como, por exemplo, o PostGreSQL.

O MySQL é desenvolvido e mantido pela empresa MySQL AB, que também oferece uma versão comercial (paga). Esse Sistema Gerenciador de Banco de Dados (SGBD) também é multi-plataforma, sendo compatível com o Windows, Linux, BSDs, entre outros sistemas operacionais. As tabelas criadas podem ter tamanho de até 4 GB. Fora isso, o MySQL é compatível com várias linguagens de programação, tais como PHP, C, Java, Visual Basic, entre outros.

Já que o MySQL não tem recursos sofisticados, para aplicações não muito complexas, ele é uma ótima indicação. O fato de utilizar somente recursos essenciais (o

---

[5] duração das ligações efetuadas: curta, média e longa

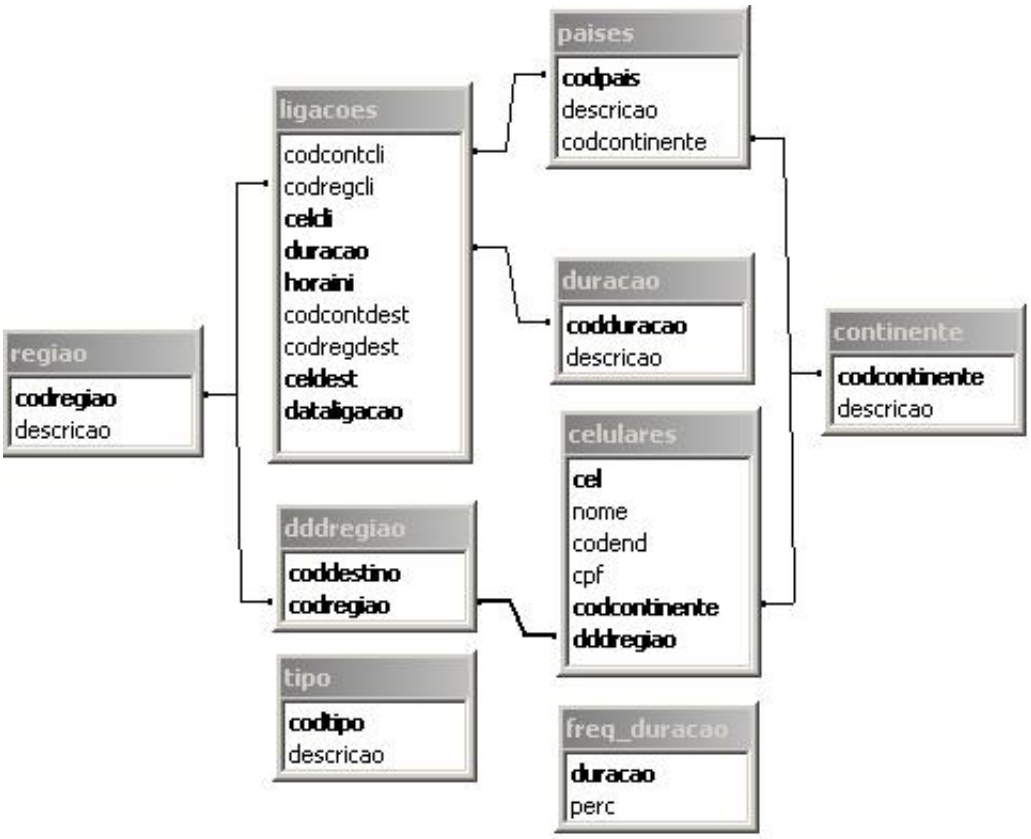
[6] local do destino da ligação: local, interurbano e internacional

que atende à maioria das aplicações de banco de dados existentes) fez com que sua adoção e conseqüentemente o suporte fosse grande. Trata-se do banco de dados de código aberto mais utilizado no mundo.

De acordo com as informações obtidas e a análise dos arquivos de ligações obtidos de uma operadora, foi implementado um banco de dados no MySQL para conter os informações das ligações efetuadas pelos usuários de celulares.

A figura 4 mostra o relacionamento entre as tabelas do BD de ligações para que se possa perceber o domínio da aplicação e os seus possíveis filtros baseados na duração e o tipo de ligação conforme o país, região ou continente.

Figura 4: Modelagem



A tabela *ligações* contém as chamadas que os clientes executam. Os atributos *codcontdest* e *codregdest* fornecem a informação que caracteriza se a ligação é internacional, interurbano ou local, métrica destino como já vimos na tabela 1.

As freqüências são armazenadas na tabela *freq_duracao* para cálculo das probabilidades *a priori* através da fórmula da figura 1 Análise de Probabilidades

O local de destino da ligação exigiu que fosse criada as tabelas *regiao* e *dddregião*, as quais contam as regiões e os respectivos códigos de área.

4.3. Aplicação

A aplicação foi desenvolvida em Delphi, um ambiente de desenvolvimento Rápido (RAD) no Windows e Linux que usa o Object Pascal como linguagem de programação. A escolha foi baseado na experiência já adquirida de vários anos de programação nesta linguagem.

O protótipo desenvolvido resulta em uma rede bayesiana, que pode ser utilizada para realizar previsões sobre o comportamento e alterações de dados usados em telefonia celular, podendo ser utilizada também por sistemas especialistas de gerência de redes.

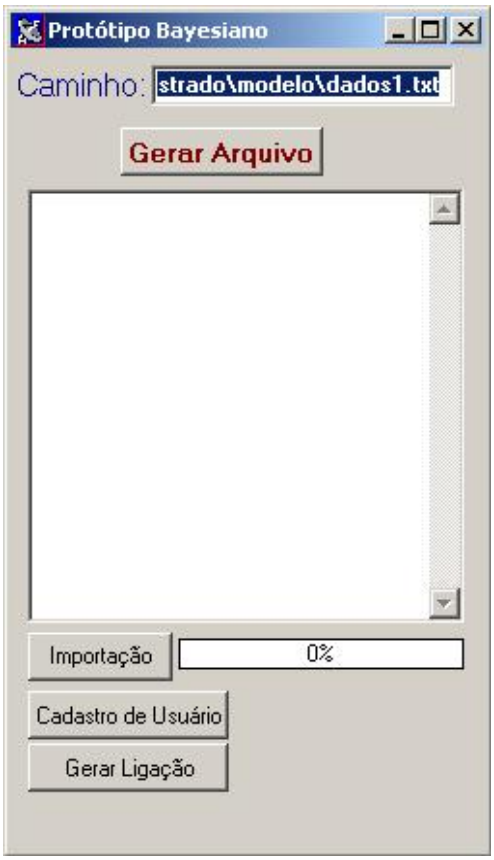
O foco esta baseado nas métricas mostradas na tabela abaixo.

Tabela 1: Métricas

Grupos	Data	Horário	Chamada	Duração
1	**Dia Útil**	**S/Desconto**	**Local**	**Curta**
2	**Dia Útil**	**S/Desconto**	**Local**	**Média**
3	**Dia Útil**	**C/Desconto**	**Local**	**Curta**
4	**Dia Útil**	**C/Desconto**	**Local**	**Média**
5	**Dia Útil**	**S/Desconto**	**Interurbana**	**Curta**
6	**Não Útil**	**C/Desconto**	**Local**	**Curta**
7	**Não Útil**	**C/Desconto**	**Local**	**Média**
8	**Não Útil**	**C/Desconto**	**Interurbana**	**Curta**
9	**Não Útil**	**C/Desconto**	**Interurbana**	**Média**

Para os dados serem importados pelo Protótipo Bayesiano de um arquivo com aproximadamente três mil ligações da operadora, deve-se clicar no botão importação, como mostra a figura 5. As ligações tornam possível o cálculo das freqüências e inserção das mesmas na tabela *freq_duracao* tornando possível a criação dos nós e a geração da Rede Bayesiana.

Figura 5 : Protótipo Bayesiano



Os usuários são cadastrados e relacionados aos números de celulares importados do arquivo de ligações podendo ser possível a identificação posterior dos mesmos para as ligações futuras. As simulações de ligações são realizadas selecionando

o cliente que deseja ligar, digitando o número com quem se quer falar e a ligação está pronta a se iniciar, como mostra a figura 6.

Figura 6 : Simulação de Ligações

Após iniciar a ligação as comparações de cálculos de duração de ligação são realizadas e o perfil do cliente, já mapeado na base de dados, serve para a comparação com a ligação iniciada e conseqüente deslocamento pelos nós da rede bayesiana. Caso a ligação saia do perfil do cliente um alarme pode ser disparado pelo sistema.

A arquitetura do Sistema implementado neste trabalho é composta das seguintes partes:

- **Estação gerente / Central -** Estação de trabalho onde é executado o programa gerente, responsável pela identificação da chamada;
- **Base de dados -** Arquivos onde serão gravados os dados coletados pela operadora;
- ***Data Mining* –** Processo responsável pela preparação dos dados coletados e pelo processo de *Data Mining* pelo Protótipo.
- **Rede bayesiana –** É composta por variáveis, seus respectivos atributos e, pela relação entre as variáveis que são estabelecidas através da definicão de regras e/ou fatos. Esta rede é representada por distribuições de probabilidades para as hipóteses diagnósticas.

- **Interface –** Tem como objetivo principal, fazer a comunicação entre o usuário e a rede bayesiana, exibe todas as informações durante as consultas.
- **Usuário –** Celular que fará a ligação.

Em cima desta arquitetura o algorítmo bayesiano fará constantes verificações na base de dados e verificará alterações no BD, assim recalculando as probabilidades e gerando novos perfis para os usuários.

4.4. Ferramentas

Existem diversas plataformas de gerência de redes no mercado como por exemplo o *NetView*, da IBM; a *Sun Net Manager* da SUN e a *HP OpenView* da Hewlett Packard - HP. Mesmo utilizando as facilidades destas plataformas, como monitoração de variáveis e visualizações gráficas, gerenciar uma rede ainda é uma tarefa difícil. A todo momento o administrador, que tem suas decisões apoiadas basicamente em dados, depara-se com situações que exigem tratamento da incerteza. Apesar dos serviços oferecidos, nenhuma das plataformas atuais é capaz de identificar problema e sugerir ações corretivas, deixando ao administrador o encargo de interpretar gráficos e valores de variáveis.

4.4.1. Nética

Netica é uma ferramenta ampla para redes Bayesianas e diagramas de influência. Pode-se construir, aprender, modificar, transformar, entre outras operações em uma interface amigável e uma máquina de inferência poderosa. Os gráficos da rede são de alta qualidade, sendo permitidas diversas configurações de visualização.

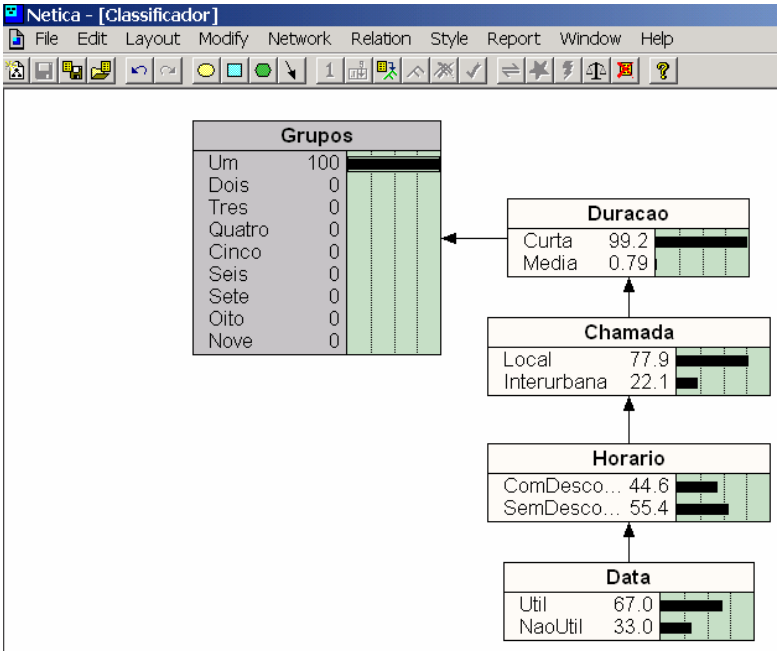
O Shell Netica, possui uma interface gráfica interativa, onde cria os nodos da Rede Bayesiana que representam as variáveis, definir os arcos entre esses nodos que são as dependências causais entre as variáveis e atribuir as probabilidades condicionais a priori para cada nodo.

Suas características principais são:

- Transforma redes de crença em árvores de estados (para prover um raciocínio mais rápido);
- Pode aprender relações probabilísticas a partir de dados;
- Permite inserir relações probabilísticas por equações, com uma biblioteca de funções probabilísticas e matemáticas fornecidas com o software. As equações podem ser determinísticas ou probabilísticas, e para variáveis discretas ou contínuas;
- Operação gráfica bastante simples e poderosa.

A Rede Bayesiana de telefonia é composta por cinco nodos, onde um representa os grupos e os restantes representam as variáveis consideráveis para se definir o perfil de ligações como mostra a figura 7. Uma vez que se tem a parte qualitativa da rede representada no modelo, pode-se construir a parte quantitativa da rede determinando as probabilidades condicionais à priori para cada variável. Esta etapa de determinação das partes Qualitativa e Quantitativa da rede compõe a fase de Aquisição de Conhecimento de um Sistema Especialista.

Figura 7: Netica



Aos fatos e as regras são associadas às incertezas presentes no domínio através de valores de probabilidades. O raciocínio realizado pelo sistema deve considerar estas probabilidades iniciais (probabilidades a priori) obtidas da pesquisa, para que o sistema, a partir dos dados de entrada (evidências), possa calcular o diagnóstico mais provável (probabilidades a posteriori).

Estatisticamente se comprava o grande número de ligações são curtas e locais, sem desconto e em dias úteis, conforme figura 7. Assim podemos classificar as futuras ligações nos nove grupos criados conforte visto na tabela 1 de métricas. A modelagem bayesiana, juntamente com o software Netica, se mostrou muito interessante para esta abordagem.

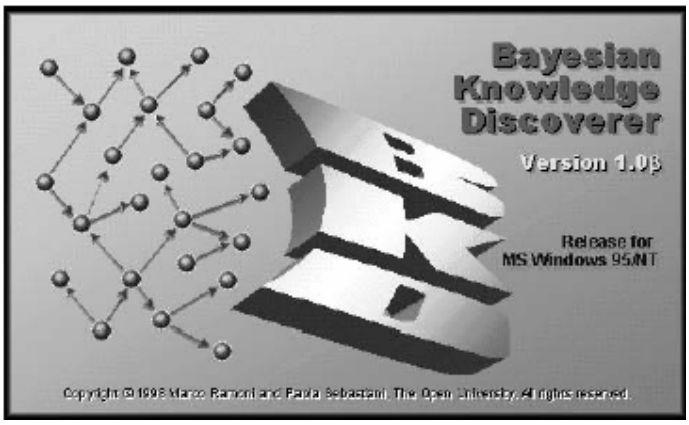
O transporte das porcentagens conseguidas pela pesquisa para o software Netica se passou sem nenhum problema, tendo em vista que os dados foram gerados de uma operadora em formato específico com geração posterior de dados padronizados pelo protótipo seguindo o padrão da amostragem.

4.4.2. Bayesian Knowledge Discoverer (BKD)

É uma ferramenta gratuita desenvolvida em commonlisp que permite a construção de redes Bayesianas, a propagação de evidências na rede e o aprendizado de novas redes a partir de dados da implementação do algoritmo Bound and Collaps.

No software, a construção da rede é feita em um editor gráfico que gera um arquivo do tipo BBN (bayesian belief network), entretanto, pode-se salvar a rede gerada no formato BNIF, assim como importar uma rede neste formato. Os dados para o aprendizado devem estar armazenados em um arquivo texto puro (ASCII). A interface gráfica permite, além de outras operações, alterações na estrutura da rede de forma simplificada.

Figura 8– tela de abertura do software BKD



Capítulo V

5. Conclusões

Aprender redes Bayesianas a partir de dados é um processo que consiste em duas etapas distintas: o aprendizado da estrutura e o aprendizado dos parâmetros numéricos da rede.

O foco deste trabalho é o aprendizado que permite recuperar a estrutura de uma rede Bayesiana que representa um banco de dados de casos sobre um determinado domínio.

Os algoritmos que constróem redes Bayesianas utilizando métodos de busca e pontuação, de um modo geral, usam métodos heurísticos de busca para construir um modelo e avaliá-lo através de um método de pontuação. É um processo contínuo que termina quando a pontuação atribuída para um determinado modelo não pode mais ser melhorada. Diversos métodos de pontuação têm sido aplicados nestes algoritmos tais como o método de pontuação com enfoque Bayesiano.

Um dos principais problemas em aprendizado de redes Bayesianas é o número de parâmetros a serem aprendidos para uma determinada variável, o qual cresce exponencialmente com o número de pais que ela possui. Obse rvamos, na literatura, que este mesmo problema surge quando a especificação destas probabilidades acontece com o auxílio de especialistas.

Quando é selecionada uma estrutura de rede, é necessário especificar os parâmetros numéricos para completar a rede Bayesiana. Estes parâmetros podem ser estimados diretamente apartir do banco de dados.

6. Trabalhos Futuros

Com base na análise do que foi proposto neste trabalho pode-se recomendar trabalhos futuros. Alguns possíveis trabalhos futuros são propostos abaixo destacando a aplicação de técnicas de amostragem de dados.

6.1. Aplicação de técnicas de amostragem de dados

Quando temos uma amostra muito grande de dados se faz necessário a aplicação de técnicas de amostragem de dados, para que o sistema possa ser aplicado a qualquer quantidade de informações

Capítulo VI

7. REFERÊNCIAS

[ABR00] ABRAÃO, Carlos. Divórcio no Brasil. Juris Brasil, São Paulo, v. 42, n. 3, jun. 2000. Disponível em: http://sites.uol.com.br/bcncb/Mapabr.htm. Acesso em: 25 jan. 2000.

[BU03] Bussab,Wilton de º, 1940 – Estatística Básica / Wilton de º Bussab, Pedro ª Morentin – 5. Ed. – São Paulo : Saraiva, 2003.

[BRA96] BRACHMAN, Ronald J.; Anand, Tej. The Process of Knowledge Discovery In Databases. In: Advances in Knowledge Discovery and Data Mining. Menlo Park: AAAI Press, 1996. 611 p. p. 37-57 TM.

[BRA00] Brandt, Sami, Heikkonen, Jukka. “A Bayesian weighting principle for the fundamental matrix estimation”. Pattern Recognition Letters 21 (2000) 1081±1092

[CAR00] Carneiro, A. Lenin & da Silva, Wagner T. Aprendizado em Redes Bayesianas: Implementação de Algoritmo. In: Anais do IBERAMIA/SBIA 2000 - Workshop on Probabilistic Reasoning in Artificial Inteligence. Atibaia, São Paulo, Brazil, November 19-22, 2000, pp 14-20.

[FAY96] FAYYAD,UsamaM.; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From Data Mining to Knowledge Discovery: An Overview. In: Advances in Knowledge Discovery and Data Mining. Menlo Park: AAAI Press,1996.611

p. p.11-34.

[FAY98] FAYAAD, Usama ; Taming the Giants and the Monsters: Mining Large Database for Nuggets of Knowledge. Microsoft Research, One Microsoft Way Redmond, WA: [s.n], 1998.

[FRA97] FRANCESCHI, ªS.M.; ROCHA, M.ª;WEBER, H.L.; WESTPHALL, C.B. "Employing Remote Monitoring and Artificial Intelligence Techniques to Develop the Proactive Network Management". Proceedings of the International Workchop on Applications of Neural Networks to Telecommunication 3. Laurence Erlbaum Associates, Publishers. Mahwah, (NJ), USA. 1997.

[FRA00] Frans M. Coetzee, Steve Lawrence, C. LeeGiles. BayesianClassification and Feature selection from finit Data Sets - Conference on Uncertainty in Artificial Intelligence (UAI 2000) , Stanford, CA, pp. 88-97, 2000.

[GUO00] Guodong Guo, Songde Ma. "Bayesian learning, global competition and unsupervised image segmentation", Pattern Recognition Letters 21 (2000) 107±116

[HAN] Han, Eui-Hong (Sam); Karypis, George; Kumar, Vipin, "Scalable Parallel Data Mining for Association Rules".

[HKS] Han, Jiavvei; Koperski, Krzysztof; Stefanovic, Nebojsa, "GeoMiner: A System Prototype for Spatial Data Mining".

[HRU98] Hruschka Jr., E.R. & da Silva, Wagner T. Apredizado e Inferência Bayesiana no Diagnóstico de Doenças Pulmonares. In: III Simpósio Nacional de Informática – III SNI, pp. 61-65. Santa Maria, RS, 22 a 25 de Setembro de 1998.

[MEN98] MENESES, Claudio J.; GRINSTEIN, Georges G. Categorization and Evaluation of Data Mining Techniques. In: Data Mining. Southampton: WIT Press, 1998. P.53-80.TM

[MOG00] Moghaddam, Baback, Jebara , Tony, Pentland, Alex. “Bayesian face recognition”. Pattern Recognition 33 (2000) 1771}1782

[PAL] Palmer ,Christopher R.; Faloutsos Christos ,“Density Biased Sampling: An Improved Method for Data Mining and Clustering”.

[PEN00] Peña, J.M., Lozano, J.A., Larrañaga, P. “An improved Bayesian structural EM algorithm for learning Bayesian networks for clustering”. Pattern Recognition Letters 21 (2000) 779±786

[PEARL88] Pearl, Judea.. Probabilistic Reasoning in Intelligent Systems. San Mateo, CA: Morgan Kaufmann. 1988.

[SEM01] SEMYR, JÈRUND GA; NATVIG, BENT; SÈRENSEN, ERIK. “A Comparison of Two Sequential Metropolis- Hastings Algorithms with Standard Simulation Techniques in Bayesian Inference in Reliability Models Involving a Generalized Gamma Distribution - Methodology and Computing in Applied Probability”, 3, 51±73, 2001.

[TZE00] Tze Fen Li. “Bayes empirical Bayes approach to unsupervised learning of parameters in pattern recognition - Pattern Recognition 33 (2000) 333}340

[VER00] Veronez, Cleverson Alessandro, “Gerência de Desempenho do Tráfego em Redes Utilizando Baseline Bayesiana”. Dissertação submetida à Universidade Federal de Santa Catarina para a obtenção do grau de Mestre em Ciência da Computação. 2000.

[WES96] WESTPHALL, C. B. & KORMANN, L. F., "Usage of the TMN Concepts for the Configuration Management of ATM Network", in Broadband Strategies and Technologies for Wide Area and Local Access Network, Germany, Berlin, Oct. 10-11, 1996, Roberto Vercelli, Editor, ISBN 0-8194-2357-2, Proc. SPIE Vol. 2953, pages 250-7 (1996).

[WES98] WESTPHAL, C. Becker, BLAXTON, T. “Data Mining Solutions – Methods and Tools for Solving Real-World Problems”. John Wiley & Sons, Inc. New York, N.Y. USA. 1998.

[YI99] Yi-Hong Tseng, His-Jian Lee - Pattern Recognition Letters 20 (1999) 791±806 Recognition-based handwritten Chinese character segmentation using a probabilistic Viterbi algorithm

ANEXO I

A seguir é listado o código da implementação do algoritmo de aprendizado K2.

```
Dim vNomesVars()
Dim vPais(), vRis(), vRisOcorr
Dim nPais, nVars As Integer
Dim nCasos As Long
Dim NomeArquivo As String
Dim NomeTabela As String
Dim TempoFim, TempoIni
Dim vLogFatRi() As Double
Sub CopiaVetor(vFonte(), vDestino())
Dim Cont, Lb, Ub As Integer
Lb = LBound(vFonte)
Ub = UBound(vFonte)
If IsEmpty(vDestino(1)) Then
ReDim vDestino(1 To UBound(vFonte))
For Cont = Lb To Ub
vDestino(Cont) = vFonte(Cont)
Next
Else
For Cont = Lb To Ub
ReDim Preserve vDestino(1 To UBound(vDestino) + 1)
vDestino(UBound(vDestino)) = vFonte(Cont)
Next
End If
End Sub
Sub UneVetor(vFonte1(), vFonte2(), vSaida())
CopiaVetor vSaida, vFonte1
CopiaVetor vSaida, vFonte2
End Sub
Sub PegaVars(nVars, vNomeVars(), nCasos)
Dim i As Integer
nVars = dados.Recordset.Fields.Count
dados.Recordset.MoveLast
nCasos = dados.Recordset.RecordCount
ReDim vNomesVars(1 To nVars)
For i = 1 To nVars
vNomesVars(i) = dados.Recordset.Fields(i - 1).Name
Next
dados.Recordset.MoveFirst
End Sub
Sub RiNos()
Dim StrSql As String
Dim Cont, Cont2 As Integer
Dim vOcorr()
ReDim vRis(1 To nVars)
ReDim vRisOcorr(1 To nVars)
ReDim vLogFatRi(1 To nVars)
For Cont = 1 To nVars
79
StrSql = "Select " + vNomesVars(Cont) + ", count(" + vNomesVars(Cont) + ") as Ocorre from " +
Combo1.Text
```

```
StrSql = StrSql + " Group by " + vNomesVars(Cont) + ";"
dados.RecordSource = StrSql
dados.Refresh
vRis(Cont) = dados.Recordset.RecordCount
vLogFatRi(Cont) = LogFat(vRis(Cont))
dados.Recordset.MoveFirst
ReDim vOcorr(1 To vRis(Cont))
For Cont2 = 1 To vRis(Cont)
vOcorr(Cont2) = dados.Recordset.Fields(1).Value
dados.Recordset.MoveNext
Next
vRisOcorr(Cont) = vOcorr
Next
End Sub
Sub K2()
Dim i, j, nMaxPais As Integer
Dim valor_g, valor_gj, delta As Double
Dim Flag As Boolean
Dim vNos()
dados.Refresh
nMaxPais = 4
PegaVars nVars, vNomesVars(), nCasos
ReDim vNos(1 To nVars)
RiNos
ProgressBar.Max = nVars
ProgressBar.Visible = True
TempoIni = Time()
For i = 2 To nVars
ProgressBar.Value = i
valor_g = calcula_g1(i)
Flag = True
nPais = 0
While Flag And nPais < nMaxPais
If nPais < i - 1 Then
seleciona_candidato i, j, valor_gj
If valor_gj > valor_g Then
valor_g = valor_gj
nPais = nPais + 1
ReDim Preserve vPais(1 To nPais)
vPais(nPais) = j
txtResultados.Text = txtResultados.Text + vNomesVars(j) + "->" + vNomesVars(i) + Chr(10)
txtResultados.Refresh
Else
Flag = False
vNos(i) = vPais
ReDim vPais(1 To 1)
End If
Else
Flag = False
80
vNos(i) = vPais
ReDim vPais(1 To 1)
End If
Wend
Next
TempoFim = Time()
MsgBox "Ok. Rede Gerada." + "Tempo = " + Str(TempoFim - TempoIni * 1)
End Sub
```

```
Sub seleciona_candidato(i, j, valor_gj)
Dim c, c1, nPaisCand As Integer
Dim g_Candidato As Double
Dim JaPai As Boolean
valor_gj = -100000 'valor inicial grande
j = 0
For c = i - 1 To 1 Step -1
'evita se já eh pai
JaPai = False
If nPais > 0 Then
For c1 = 1 To nPais
If c = vPais(c1) Then
JaPai = True
End If
Next
End If
'calcula o valor da funcao g para este candidato
If Not JaPai Then
g_Candidato = g(i, c)
If g_Candidato > valor_gj Then
valor_gj = g_Candidato
j = c
End If
End If
Next
End Sub
Function g(ByVal i As Integer, ByVal c As Integer) As Double
Dim TextSql As String
Dim Qi, Nijk, Nij, Ri, k, j As Integer
Dim calcG, Parcial, LogFatRi As Double
Dim ConfPais(), vCand()
If nPais > 0 Then
ReDim vCand(1 To nPais)
CopiaVetor vPais, vCand
ReDim Preserve vCand(1 To nPais + 1)
vCand(nPais + 1) = c
Else
ReDim vCand(1 To 1)
vCand(1) = c
End If
Ri = vRis(i)
CalculaQi i, ConfPais(), vCand
81
Qi = UBound(ConfPais)
Parcial = 0
calcG = 0
LogFatRi = vLogFatRi(i)
For j = 1 To Qi
Nij = 0
Nijk = 0
SumNijk = 0
'calcula Nij e ultimo termo
For k = 1 To Ri
Nijk = calculaNijk(i, k, ConfPais, j, vCand)
Nij = Nij + Nijk
SumNijk = SumNijk + LogFat(Nijk)
Next
Parcial = LogFatRi - LogFat(Nij + Ri - 1) + SumNijk
```

```
calcG = calcG + Parcial
Next
g = calcG
End Function
Function LogFat(X) As Double
Dim Soma As Double
Dim i As Integer
If X = 0 Then
Soma = 0
Else
For i = 1 To X
Soma = Soma + (Log(i) / Log(10#))
Next
End If
LogFat = Soma
End Function
Sub CalculaQi(ByVal i As Integer, ConfPais(), vCand())
Dim StrSql As String
Dim nPaisCand As Integer
Dim Cont, Cont2 As Integer
nPaisCand = UBound(vCand)
StrSql = "Select "
For Cont = 1 To nPaisCand
StrSql = StrSql + vNomesVars(vCand(Cont)) + ","
Next
StrSql = StrSql + "count(" + vNomesVars(i) + ")"
StrSql = StrSql + " from " + Combo1.Text + " Group by "
For Cont = 1 To nPaisCand
StrSql = StrSql + vNomesVars(vCand(Cont))
If Cont < nPaisCand Then
StrSql = StrSql + ","
Else
StrSql = StrSql + ";"
End If
Next
dados.RecordSource = StrSql
82
dados.Refresh
If Not (dados.Recordset.EOF And dados.Recordset.BOF) Then
dados.Recordset.MoveLast
End If
nQi = dados.Recordset.RecordCount
dados.Recordset.MoveFirst
ReDim ConfPais(1 To nQi)
For Cont = 1 To nQi
ReDim vTemp(1 To nPaisCand)
For Cont2 = 1 To nPaisCand
vTemp(Cont2) = dados.Recordset.Fields(Cont2 - 1).Value
Next
ConfPais(Cont) = vTemp
dados.Recordset.MoveNext
Next
End Sub
Function calculaNijk(i, k, ConfPais, j, vCand()) As Integer
Dim StrSql As String
Dim Cont As Integer
Dim CPais()
CPais = ConfPais(j)
```

```
StrSql = "Select "
For Cont = 1 To UBound(vCand)
StrSql = StrSql + vNomesVars(vCand(Cont)) + ","
Next
StrSql = StrSql + vNomesVars(i) + " from " + Combo1.Text + " Where "
For Cont = 1 To UBound(vCand)
StrSql = StrSql + vNomesVars(vCand(Cont))
StrSql = StrSql + "='" + Trim(Str(CPais(Cont))) + "' and "
Next
StrSql = StrSql + vNomesVars(i) + "='" + Trim(Str(k)) + "';"
dados.RecordSource = StrSql
dados.Refresh
If Not (dados.Recordset.EOF And dados.Recordset.BOF) Then
dados.Recordset.MoveLast
End If
calculaNijk = dados.Recordset.RecordCount
End Function
Private Sub cmdBanco_Click()
Dialog.Show Modal
End Sub
Private Sub cmdIniciar_Click()
K2
End Sub
Private Sub cmdSair_Click()
Unload Me
End Sub
Function calcula_g1(ByVal i As Integer) As Double
Dim Ri1 As Integer
Dim Nij1, Nijk1 As Double
Dim StrSql As String
Dim Cont As Integer
Ri1 = vRis(i)
Nij1 = LogFat(nCasos + Ri1 - 1)
For Cont = 1 To Ri1
Nijk1 = Nijk1 + LogFat(vRisOcorr(i)(Cont))
Next
calcula_g1 = LogFat(Ri1 - 1) - Nij1 + Nijk1
End Function
'Formulário auxiliar para escolher o banco de dados e a tabela (Dialog)
Option Explicit
Dim NomeBanco As String
Dim CaminhoNome As String
Dim X As Integer
Private Sub Combo1_Change()
NomeBanco = Combo1.Text
End Sub
Private Sub Form_Activate()
cdgEscolheBco.ShowOpen
CaminhoNome = cdgEscolheBco.FileName
Data1.DatabaseName = CaminhoNome
Data1.Refresh
For X = 0 To Data1.Database.TableDefs.Count - 1
Combo1.AddItem Data1.Database.TableDefs(X).Name
Next
Combo1.SetFocus
End Sub
Private Sub Form_Load()
Combo1.Clear
```

```
End Sub
Private Sub OKButton_Click()
frmPrincipal.Combo1.Text = Combo1.Text
frmPrincipal.dados.DatabaseName = CaminhoNome
frmPrincipal.dados.RecordSource = Combo1.Text
frmPrincipal.cmdIniciar.Enabled = True
Unload Me
End Sub
```